Google Trends peut-il aider à prévoir le chômage ?

Diane THIERRY

Connaissez-vous les voitures intelligentes ? Les gants Sing-IO qui changent le langage des signes en discours audio ? L'IA miniature ou encore les Google Glasses ? Chacune de ces inventions reflète le monde connecté dans lequel nous vivons depuis quelques années. De telles avancées ont été possibles au moyen de l'essor des nouvelles technologies qui sont venues changer nos modes de vie, de consommation et nos habitudes en profondeur. Considérée comme la troisième révolution industrielle, l'ère de la technologie s'accompagne d'une transformation du numérique où chaque jour environ 2,5 trillions d'octets de données sont créés. Certaines plateformes ont profité de cet afflux de données pour utiliser les informations qu'elles contiennent, et les mettre à disposition des utilisateurs. Il est un nouvel outil depuis 2006 qui informe des tendances du web en donnant la popularité relative d'un ou plusieurs mots recherchés en ligne ; c'est **Google Trends** (GT). Par ses nombreux avantages, l'outil a suscité un intérêt croissant chez les conjoncturistes qui savent largement tirer parti de ces informations pour améliorer et préciser leurs prévisions.

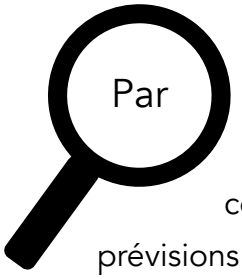
• *Google Trends, qu'est ce que c'est ?*

L'outil né en 2006 a pour but d'identifier la popularité de certains termes dans les recherches internet des individus, à une période donnée. Les valeurs des fréquences de recherche qu'il met à disposition sont calibrées entre 0 et 100, où 100 correspond à la recherche la plus importante de la période ou de la région. L'outil a notamment servi dans l'analyse des comportements des français lors des 2 mois de confinement : on a alors vu que l'intérêt pour des solutions aux insomnies ou pour des informations liées aux masques est monté en flèche, et que l'association "coronavirus" et "complot" dans les recherches a battu des records la veille du début du confinement.[1] Par conséquent, l'outil Google Trends offre la possibilité de connaître en temps réel les tendances de recherches qui traduisent à la fois les goûts, les envies, mais aussi les préférences des internautes.

Principaux avantages :

- ✓ Gratuit
- ✓ Facile d'accès
- ✓ Disponible en Open Data
- ✓ Données ajustées [0:100]
- ✓ Mis à jour quotidiennement
- ✓ Disponible pour tous les pays
- ✓ Représentatif des recherches internet*

** Sur les 53,1 millions d'internautes en France en 2019, 95% d'entre eux effectuent leurs recherches sur le moteur Google.[2]*



ses nombreux avantages, l'outil a fait déjà l'objet de multiples études quant à son apport pour améliorer les techniques de prévisions. Dans ce travail nous cherchons à savoir s'il peut être un bon indicateur des prévisions du taux de chômage. Par rapport aux travaux déjà réalisés sur le sujet, nous allons apporter une contribution non des moindres, puisqu'il s'agit de vérifier la nature de la relation de toutes les variables explicatives au taux de chômage, avant de les inclure aux différents modèles. En effet, le meilleur moyen pour connaître l'apport de GT dans la prédiction du chômage, est de construire plusieurs modèles en incluant ou pas cette nouvelle variable Google, puis réaliser différentes prévisions que nous comparerons par la suite. Si les prévisions du modèle incluant GT sont meilleures que celles d'un modèle sans cette dernière, alors nous pourrons conclure positivement quant à sa contribution pour prédire le chômage. Les modèles avec variables explicatives que nous construisons sont donc composés uniquement de variables ayant une relation stable et robuste avec Y_t. Nous disposons pour cette étude de 6 variables équidistantes : le taux de chômage qui constitue la variable à expliquer, 1 variable issue de l'outil Google Trends (les recherches du mot 'emploi') et 4 variables macroéconomiques traditionnelles : le taux d'intérêt, la production industrielle, la population active et les effectifs dans l'industrie manufacturière.

• Méthodologie statistique

Nous réalisons dans ce dossier une double analyse, à la fois par des modélisations simples, puis par des modélisations plus complexes puisque composées de variables exogènes. Pour les modélisations simples nous appliquons la méthode de Box-Jenkins pour trouver le meilleur modèle ARIMA, méthode qui consiste en 4 étapes distinctes : stationnarisation de la série, identification du modèle, estimation, et vérification des résidus. Nous réalisons les modélisations plus complexes à l'aide de modèles ARX dont les principaux avantages sont la linéarité des estimations, le traitement de diverses variables et le faible nombre de paramètres à identifier. Comme nous l'avons dit précédemment nous n'intégrerons dans ces modèles que les variables dont la relation avec Y_t a été vérifiée. Pour ce faire, nous regarderons d'une part si elles sont intégrées du même ordre, et d'autre part si la relation qui les lie n'est pas fallacieuse. Ainsi, cette double modélisation permet une analyse assez complète du sujet; la comparaison des prévisions issues des différents modèles va montrer lesquelles sont les plus précises et donc quel modèle est le plus pertinent.

• *Quels avantages des méthodes par rapport à l'étude ?*

⟶ **Modélisations ARIMA** : Basées sur les valeurs passées de la variable à expliquer, la méthode ARIMA modélise les variations de Y_t de manière simple avec peu de paramètres, et peut offrir de bons résultats d'estimation pour une faible complexité. De même, les modèles issus de telles estimations sont facilement interprétables et l'on peut aisément prévoir les variations future de la série par un processus ARIMA.

⟶ **Modélisations ARX** : La méthode directe d'identification en boucle fermée des modèles ARX présente de nombreux avantages. Combinant à la fois des entrées v_i de variables exogènes et des sorties de y_i, les modèles ARX ou ARMAX offrent de bons résultats. Les paramètres y sont déterminés par la méthode des moindres carrés ou la matrice instrumentale.

⟶ **Apport de ces méthodes par rapport à l'étude** : Cette double modélisation permet de quantifier réellement l'impact de l'ajout de variables explicatives dans la qualité des prévisions. La simplicité et la pertinence de ces méthodes nous permettront de mesurer la contribution de Google Tendances pour préciser les prévisions de taux de chômage en France.

• Résultats obtenus

La cointégration des variables dans le but de les inclure aux modèles ARX s'est révélée concluante pour 3 des 5 variables explicatives. Les variables "popularité du mot 'emploi'", "production industrielle" et "effectifs dans l'industrie manufacturière" ont une relation pertinente avec le taux de chômage. Cependant, après avoir regardé la matrice de corrélation des variables différenciées, nous nous sommes rendu compte qu'il n'existait aucune relation entre le taux de chômage et les effectifs dans l'industrie manufacturière (CC=0), tandis que sur les variables en niveau le coefficient était de 0,29. Nous avons donc créé 3 modèles ARX, l'un composé des 2 variables cointégrées ensemble et les deux autres avec chacune d'elles prisent individuellement. De même, nous avons modélisé un processus AR[1] à partir de la méthodologie ARIMA en suivant l'approche de Box-Jenkins.

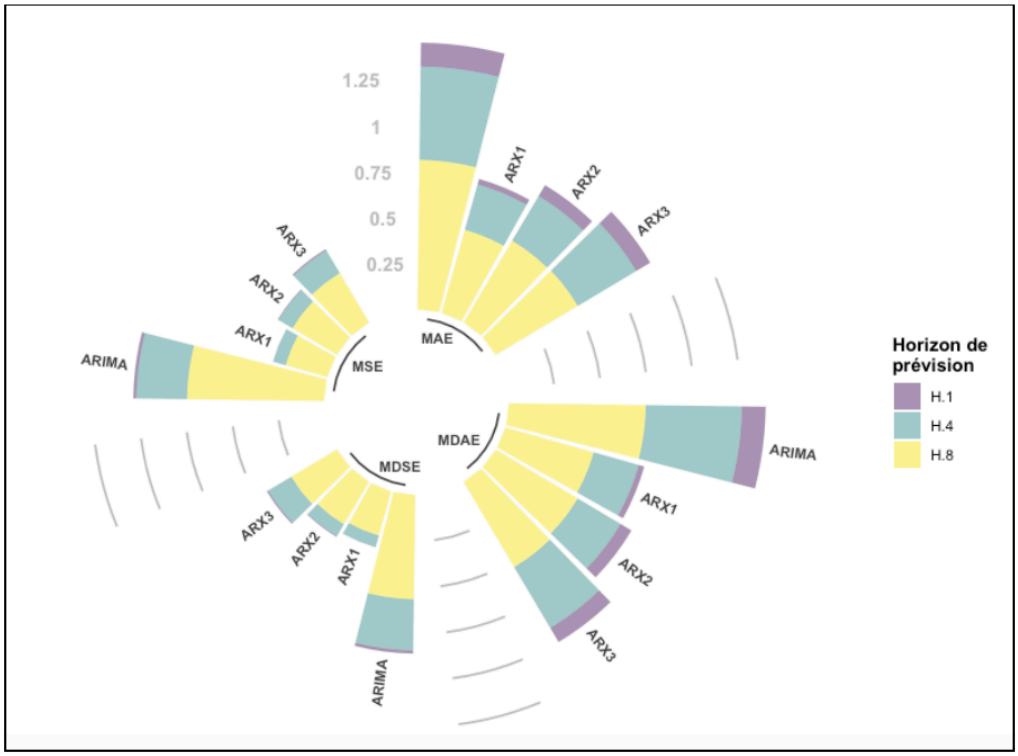
Tableau n°1 : Récapitulatif des modèles estimés

	ARIMA	ARX1	ARX2	ARX3
Variables explicatives	/	- Google Trends - Production industrielle	Google Trends	Production industrielle
Nombre de coefficients significatifs à 10%	1/1	1/2	1/2	2/2
Écart-type de régression	0,24	0,20	0,22	0,21
R^2	0,06	0,38	0,20	0,29
Log de vraisemblance	1,59	13,83	6,11	9,85

Les principales caractéristiques des modèles estimés figurent dans le tableau n°1, on voit que le meilleur modèle en termes de qualité est celui qui est composé de 2 variables exogènes. Il explique 38% des variations du taux de chômage quand les autres n'en expliquent pas plus de 29, il a la vraisemblance la plus forte et l'écart-type le plus faible. Le modèle ARIMA a contrario, possède les coefficients de vraisemblance et de détermination les plus faibles de tous, ainsi qu'un écart-type le plus élevé, et n'explique que 6% du taux de chômage. En utilisant ces 4 modèles nous avons pu faire les prévisions du taux de chômage à 3 horizons différents : le très court terme soit 1 trimestre, le moyen terme soit 4 trimestres, et le long terme soit 8 trimestres. Il ressort de la confrontation des différentes prévisions réalisées, que le modèle surpassant tous les autres en termes de capacité prévisionnelle est celui qui combine à la fois un indicateur macroéconomique (la production industrielle), et la variable issue de Google Trends qui vient apporter un complément d'informations qui n'est pas forcément contenu dans les variables traditionnelles. Effectivement, ce modèle est plus précis que celui ne contenant que la variable macroéconomique. Par ailleurs, si l'on compare les modèles composés d'une seule variable explicative (ARX2 : Google Trends, et ARX3 production industrielle), il apparaît en termes de taux d'erreur, que le modèle ARX2 est meilleur que ARX3, en revanche si l'on regarde le test DM ce dernier surpasse le modèle ARX2. De plus, le test de Mariano et Preve appliqué à horizon de 4 périodes dévoile que le modèle contenant la variable de Google Tendances seule, est le meilleur de tous, mais c'est le dernier à un horizon de 8 périodes. Les meilleures prévisions réalisées sont donc issues du modèle ARX1 : elles ont les taux d'erreur les plus faibles des 4 modèles estimés, comme visible sur le graphique ci-dessus.

Dans ce dossier nous nous contentons de réaliser des prévisions puisque les prédictions sont issues de lois scientifiques assurées.

GRAPHIQUE N°1 : Barplot circulaire des erreurs de prévision



• Conclusion

De plus en plus présent dans notre société, le Big Data est devenu un facteur clé pour obtenir des informations pouvant aider l'analyse économique. Dans cette étude, nous avons montré comment l'outil Google Trends permet d'améliorer et de préciser les prévisions du chômage et de l'emploi, par la modélisation et la prévision du taux de chômage français de 2004 à nos jours. En effet, le modèle combinant une variable macroéconomique qu'est la production industrielle, couramment utilisée dans la prédiction du taux de chômage, à la popularité du mot 'emploi' sous google Trends, surpasse tout autre modèle en termes de prévision. En outre, l'outil est surtout efficace lorsqu'il est combiné aux variables macroéconomiques traditionnelles - comme nous l'avons expérimenté dans le modèle ARX1. Sachant qu'il faut 6 mois environ pour trouver un travail ou en changer, les personnes cherchent un emploi en ligne bien avant d'être au chômage ou licenciées. De cette manière, l'outil GT anticipe les mouvements du marché du travail en proposant la popularité des mots relatifs à l'emploi, d'où sa grande utilité.

La période étudiée s'étendant de 2004 à nos jours, elle prend en compte la crise des Subprimes qui est venue bouleverser l'équilibre économique et donc créer des évolutions de grande variance qu'il est difficile d'analyser par la suite. Ainsi, l'atypicité de la période étudiée a entraîné des fluctuations inhabituelles des indicateurs, ce qui explique finalement que seules 2 des 5 variables explicatives ont pu être intégrées dans les modèles ARX. Pour répondre à cette limite, il pourrait être intéressant de construire des modèles avant et après crise pour voir si les conclusions sont différentes.

• *Bibliographie*

Inwin - Digital Expert, "Un outil puissant et gratuit mais méconnu Google Trends", 11/03/2019. https://www.inwin.fr/un-outil-puissant-et-gratuit-mais-meconnu-google-trends/

Le Monde, « Masques, insomnie, farine ou jogging… Ce que nos recherches Google disent du confinement », 09/05/2020. https://www.lemonde.fr/pixels/article/2020/05/09/masques-insomnie-farine-ou-jogging-ce-que-disent-nos-recherches-google-du-confinement_6039123_4408996.html